The InFile project: a crosslingual filtering systems evaluation campaign

Romaric Besançon[*], Stéphane Chaudiron[**], Djamel Mostefa[+], Ismaïl Timimi[**], Khalid Choukri[+]

[*]CEA LIST
18, route du panorama
BP 6 – 92265 Fontenay aux Roses

[**]Université de Lille 3 – GERiiCO
Domaine universitaire du Pont de Bois
BP 60149 – 59653 Villeneuve d'Ascq cedex

[+]ELDA
55-57, rue Brillat Savarin
75013 Paris

E-mail: romaric.besancon@cea.fr, stephane.chaudiron@univ-lille3.fr, mostefa@elda.org, ismail.timimi@univ-lille3.fr, choukri@elda.org

Abstract

The InFile project (INformation, FILtering, Evaluation) is a cross-language adaptive filtering evaluation campaign, sponsored by the French National Research Agency. The campaign is organized by the CEA LIST, ELDA and the University of Lille3-GERiiCO. It has an international scope as it is a pilot track of the CLEF 2008 campaigns. The corpus is built from a collection of about 1,4 millions newswires (10 GB) in three languages, Arabic, English and French provided by Agence France Press (AFP) and selected from a 3 years period. The profiles corpus is made of 50 profiles from which 30 concern general news and events (national and international affairs, politics, sports…) and 20 concern scientific and technical subjects.

1. Introduction

The InFile evaluation campaign measures the ability of filtering systems to successfully separate relevant and non-relevant documents in an incoming stream of textual information with respect to a given profile. Following Belkin and Croft (Belkin, 1992), an information filtering system is a system designed to manage unstructured or semistructured data. Information filtering systems deal primarily with textual information, involve large amounts of data incoming through permanent streams such as newswire services. Filtering is based on individual or group information profiles which assume to represent consistent and long-term information needs. From the user point of view, the filtering process is usually meant to extract relevant data from the data streams, according to the defined by the user profiles.

Information filtering systems may be exploited in different business contexts of use : for example, text routing which involves sending relevant incoming data to individuals or specific groups, categorization process which aims at attaching one or more predefined categories to incoming documents, or anti-spamming which tries to remove « junk » e-mails from the incoming e-mails.

In the InFile project, we consider the context of competitive intelligence in which the information filtering is a very specific subtask of the information management process (Bouthillier, 2003). In this approach, the information filtering task is very similar to Selective Dissemination of Information (SDI), one of the original and usual function assumed by documentalists and, more recently, by other information intermediaries such as technological watchers or business intelligence professionals.

Therefore the project will pay, during the design of the campaign protocol, a particular attention to the context of use of filtering systems by real professional users. Even if the campaign is mainly a technological oriented evaluation process, we adapt the protocol and the metrics, as close as possible, to how a normal user would proceed, including through some interaction and adaptation of his system.

Previous evaluation campaigns have been proposed in the past years on Adaptive Filtering systems, including the Text Retrieval conference (TREC) Adaptive Filtering tracks from 2000 to 2002 (Roberston, 2002) and the Topic Detection and Tracking (TDT) campaigns from 1998 to 2004 (Fiscus, 2004). The specific features of the InFile campaign compared to these previous works are presented in the following sections.

2. Goals and features of the InFile campaign

The project targets three objectives:

- First and mainly, it is an evaluation campaign involving academic and industrial participants on a crosslingual filtering task in order to compare the systems abilities and work out state of the art.
- Secondly, it is an attempt to better understand and model the human information filtering process and possibly to modelize it in evaluation protocols and metrics. In this way, we try to respect as far as possible the "ground truth" in building the set of filtering profiles, building the set of relevant documents and adapting the protocol and the metrics to this particular context.
- The third goal is to build a test collection which consists of a large documents set in three languages, a set of filtering topics and the set of the corresponding relevant documents. This evaluation package will be made widely available for the research community.

The InFile campaign can mainly be seen as a cross-lingual pursuit of the TREC11 Adaptive Filtering task, with a particular interest in the correspondence of the protocol with the ground truth of competitive intelligence professionals.

In the TDT campaigns, focus was mainly on topics defined as "events", with a fine granularity level, and often temporally restricted, whereas in InFile (similar to TREC11), topics are of long term interest and supposed to be stable, which can induce different techniques, even if some studies show that some models can be efficiently trained to have good performance on both tasks (Yang 2005).

The main features of the InFile evaluation campaign are summarized here:

- Crosslingual : English, French and Arabic are concerned by the process but participants may be evaluated on mono or bilingual runs.
- A newswire corpus provided by the Agence France Presse (AFP) and covering recent years.
- The topic set is composed of two different kinds of profiles, one concerning general news and events, and a second one on scientific and technological subjects.
- The evaluation task is performed using an automatic interrogation of participating systems with a simulated user feedback.
- Systems are allowed to use the feedback at any time to increase performance.
- Systems provide a boolean decision for each document according to each profile.
- Relevance judgments are mainly performed by human assessors.
- Participants are asked to fill a form to specify the languages used, the fields used in the profiles, and a summary of the technology used.

3. Test collections

3.1 The AFP Corpus

The InFile corpus is provided by the Agence France Presse (AFP) for research purpose. AFP is the oldest news agency in the world, and one of the three largest with Associated Press and Reuters. Although AFP is the largest French news agency, it transmits news in other languages such as English, Arabic, Spanish, German, and Portuguese.

For InFile, we selected 3 languages, (Arabic, English and French) and a 3 years period (2004-2006) which represents a collection of about one and half millions newswires for around 10 GB. Newswires are available in three languages, Arabic, English and French but are not necessarily translations from a language to another. The amount of news per year and per language is described in Table 1.

Wire	**2004**	**2005**	**2006**	***Total***
ARA	85k	81k	87k	254k
FRE	154k	139k	154k	448k
ENG	268k	245k	244k	758k
Total	508k	467k	486k	**1 462k**

Table 1 Statistics on the AFP corpus

In the InFile campaign, only 100 000 documents of each language are used for the filtering test, in order to cope with the time constraints of an interactive filtering process as described in section 4.1. These documents correspond to the set of relevant documents for the pofiles (selected as described in section 3.3) completed by a set of non-relevant documents.

News articles are encoded in XML format and follow the News Markup Language (NewsML) specifications [1]. NewsML is an XML standard designed to provide a media-independent, structural framework for multi-media news. NewsML was developed by the International Press Telecommunications Council.

3.2 Set of filtering profiles

A set of 50 profiles is prepared covering two different categories: the first group deals with general news and events concerning national and international affairs, sports, politics… and the second one deals with scientific and technological subjects. In order to be as close as possible to the "ground truth", profiles are constructed by competitive intelligence professionals from INIST[2] (the French Institute for Scientific and Technical Information

[1] http://www.newsml.org/
[2] http://international.inist.fr/

Center), ARIST Nord Pas de Calais[3] (Agence Régionale d'Information Stratégique et Technologique), Digiport[4], ONERA[5] and OTO Research[6]. Thirty of these are general profiles and twenty are scientific profiles. The practitioners constructed both the English and the French versions of the profiles while the Arabic version is translated by native speakers.

Profiles are defined with the following structure based on real existing profiles used by Competitive Intelligence (CI) professionals:

- a unique identifier,
- a title (6 words max.),
- a description (20 words max.),
- a narrative (60 words max.),
- up to 5 keywords,
- a example of relevant text (120 words max.).

Each record of the structure may have been translated by the profile writer with the exception of the samples which need to be extracted from real documents, always in order to fit with the "ground truth". This constraint is given to avoid terminological bias in the filtering process.

3.3 The relevant set of documents

The relevant set of documents is built trough two phases, a pre-submission phase and a post-submission phase of judgements. To be as close as possible to the "ground truth" and because feedback must be sent immediately after each submission, a pooling methodology is not available. So, evaluation is mainly based on a set of relevant documents provided by human experts but, at the end of the run, a limited control (via limited pooling) is performed on documents considered as relevant by at least two systems for each specific profile.

In order to provide the necessary relevance judgments, extensive searches using different retrieval systems are conducted at ELDA after the elaboration of the profiles. In this pre-submission phase, both the professional involved in the definition of the profiles and other assessors made relevance judgments on the outputs of the systems.

In a post-submission phase, additional relevance judgments are planned to be made by the assessors after submission of results by the participants, on the documents taken from the pooled submissions for each profile. It allows to identify additional relevant documents that could have been not found by the assessors at the previous stage. This control allows the organizers to eventually adjust the set of relevant documents, to improve the reliability of the feedback given to the participants during the run and to check the performance measures. If few modifications are needed, the way each system uses the feedback to increase performance can be considered as representative. If this limited pooling control detects that many modifications are needed, the results of using feedback are less reliable. All results are based on the full set of relevant documents.

4. Description of the protocol

4.1 The Evaluation process

The protocol of the InFile campaign is designed to be a realist task for a filtering system. In particular, the idea is to avoid making the whole corpus available to the participants before the campaign, but to make it available one document at a time, simulating the behavior of the newswire service. The protocol then forces participant systems to be evaluated in a one-pass test.

The protocol is interactive, and evaluation works as follows:

- the participant system connect to a server from which its gets a run identifier: if a participant wants to submit several runs, the system must connect several times to get different run identifiers;
- the system retrieves one document;
- the system filters the document, i.e. it associates the document with one or several profiles, or discard it;
- for adaptive systems, a relevance feedback can be provided for filtered documents;
- the system can retrieve a new document: a new document can only be retrieved when the previous document has been filtered.

A simulated relevance feedback is provided for adaptive systems: the idea is again to have a simulation of a realist behavior of the CI professional. In a real process, the CI professional receives the documents found relevant to a profile in a corresponding mailbox or directory, and he can read the document and decide to remove it if it was a filtering error.

In the InFile automated process, it is also the only feedback authorized: relevance feedback can only be asked on a document associated with a profile by the system, there is no relevance feedback on discarded documents.

Furthermore, we assume that a CI professional would not have an infinite patience: feedback is then limited to a given number of documents. This number has been fixed to 50, from the advice taken from real CI professionals. This tends to give more interest to systems with quick adaptivity, than to systems that needs a large amount of data to be trained, but it seemed right for the organizers to put systems in a the context of a realistic task.

A client-server architecture has been designed to handle the interactive communication between the participant systems and the InFile document server, that uses HTTP port through a Web Services architecture, in order to deal with the potential problem of the corporate firewalls of the participants' filtering systems.

[3] http://www.aristnpdc.org/
[4] http://www.digiport.org
[5] http://www.onera.fr
[6] http://www.otoresearch.fr/

Since the evaluation is done in a one-pass test, a dry run has been organized to check the technical viability of the protocol. This dry run proposes two profiles, and 50 documents to filter out. The profiles and documents samples are made available little time before the evaluation, to allow participants to adapt their systems to the format of profiles and documents, and the general information about the domains and the type of content of the documents.

The profile is the only information available for the systems at first. No positive documents examples are given for training, except for the document sample in the profile definition.

4.2 Metrics

The results returned by the participants are binary decisions on the association of a document with a profile. The results, for a given profile, can then be summarized in a contingency table of the form:

	Relevant	Not Relevant
Retrieved	a	b
Not Retrieved	c	d

On these data, a set of standard evaluation measures is computed:

- Precision, defined as $P = a/a+b$
- Recall , defined as $R = a/a+c$
- F-measure, which is a standard combination of precision and recall (Van Rijsbergen, 1979), and depends on a parameter α, defined as

$$F = \frac{(1+\alpha)*PR}{\alpha*P+R}$$

(typically, with α=1, the same importance is given to precision and recall and F-measure is the harmonic mean of the two values).

Following the TREC Filtering tracks (Hull,1999) (Robertson,2002) and the TDT 2004 Adaptive tracking task (Fiscus, 2004), we also consider the *linear utility*, defined as $u = w_1 \times a - w_2 \times b$, where w_1 is the importance given to a relevant document retrieved and w_2 is the cost of an irrelevant document retrieved.

Filtering according to linear utility is similar to filtering by estimated probability of relevance. With w_1=2 and w_2=1, it corresponds to the rule: retrieve if P(rel)>0.33

Linear utility is bounded positively (to 1 for a perfect filtering), but unbounded negatively (negative values depend on the number of relevant documents for a profile). Hence, the average value on all profiles would give too much importance to the few profiles on which a systems would perform poorly. To be able to average the value, the measure is scaled as follows:

$$u_n = \frac{\max(u/u_{\max}, u_{\min}) - u_{\min}}{1 - u_{\min}}$$

where u_{max} is the maximum value of the utility and u_{min} a parameter considered to be the minimum utility value under which a user would not even consider the following documents for the profile.

From the Topic Detection and Tracking campaigns (TDT2, 1998), other measures are also considered:

- The estimated probability of missing a relevant document, defined as $P_{miss} = c/a+c$
- The estimated probability of raising a false alarm on a non-relevant document defined as $P_{false} = b/b+d$
- The detection cost, defined as

$$c_{\det} = c_{miss} * P_{miss} * P_{topic} + c_{false} * P_{false} * (1 - P_{topic})$$

where

- c_{miss} if the cost of a missed document
- c_{false} is the cost of a false alarm
- P_{topic} is the a priori probability that a document is relevant to a given profile.

To compute average scores, the values are first computed for each profile and then averaged. Another way of averaging would be to sum up the values for all profiles in each cell of the contingency table and compute the scores on the resulting table. The first method is preferred because it allows equalizing the contribution of the profiles, whose differences are supposed to be the main source of variance in measures.

In order to measure the adaptivity of the systems, the measures are also computed at different times in the process (e.g. each 10 000 documents), and an evolution curve of the different values across time is proposed.

Additionally, two following experimental measures are used. The first one is an *originality* measure, defined as a comparative measure corresponding to the number of relevant documents the system uniquely retrieves (among participants). It gives more importance to systems that use innovative and promising technologies that retrieve "difficult" documents.

The second one is an *anticipation* measure, designed to give more interest to systems that can find the first document in a given profile. This measure is motivated in competitive intelligence by the interest of being at the cutting edge of a domain, and not missing the first information to be reactive. It is measured by the inverse rank of the first relevant document detected (in the list of the documents), averaged on all profiles. The measure is similar to the mean reciprocal rank (MRR) used for instance in Question Answering Evaluation (Voorhees, 1999), but is not computed on the ranked list of retrieved documents but on the chronological list of the relevant documents.

5. Conclusion

At this time, the InFile campaign is not achieved and we cannot present the results from the test but other significant results have already been reached: a large

corpus of structured newswires in Arabic, English and French, a set of structured profiles and a set of human validated relevant documents for the corresponding profiles.

Two works in progress still remain: the real test which will be completed by the end of June or beginning of July and the modeling of the filtering task assumed by the CI practitioners. This last work is a long term issue which will not be achieved within the InFile project.

6. Acknowledgements

The InFile project is partially funded by the *Agence Nationale de la Recherche* (ANR-06-MDCA-011-01). We also thank all the CI professionals from INIST, ARIST, Digiport, ONERA and OTO Research company who constructed the profiles and also Christian Fluhr who helps us to design the project.

7. References

Bouthillier F., Shearer K. (2003). *Assessing Competitive Intelligence Software : A Guide to Evaluating CI Technology*. Medford, Information Today Inc.

Belkin N., Croft B. (1992). Information filtering and information retrieval : two sides of the same coin. In *Communications of the ACM*, vol. 35, n°12, pp. 29-38.

Fiscus, J.G, Wheatley, B (2004) Overview of the TDT 2004 evaluation and results, In TDT'02, NIST.

Hull D., Roberston S. (1999) The TREC-8 Filtering Track Final Report, in *Proceedings of the Eighth Text REtrieval Conference* (TREC-8)

Robertson S., Soboroff I. (2002). The TREC 2002 Filtering Track Report. In *Proceedings of The Eleventh Text Retrieval Conference (TREC 2002)*. NIST Special Publication : 500-251, http://trec.nist.gov/pubs/trec11/t11_proceedings.html

Soboroff I., Robertson S. (2002). Building a Filtering Test Collection for TREC 2002. In *Proceedings of The Eleventh Text Retrieval Conference (TREC 2002)*. NIST Special Publication : 500-251, http://trec.nist.gov/pubs/trec11/t11_proceedings.html

TDT2 (1998) The Topic Detection and Tracking Phase 2 (TDT2) Evaluation Plan, NIST http://www.nist.gov/speech/tests/tdt/1998/doc/tdt2.eval.plan.98.v3.7.pdf

Van Rijsbergen, C.J. (1979) *Information Retrieval*. Butterworths, London.

Voorhees, E.M (1999) The TREC-8 Question Answering Track Report, in *Proceedings of the Eighth Text REtrieval Conference* (TREC-8)

Yang Y., Yoo S., Zhang J., Kisiel B. (2005) Robustness of adaptive filtering methods in a cross-benchmark evaluation, In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, Salvador, Brazil, pp. 98 - 105